Arthritis Australia 2024 National Research Program
Project Grant

Final report

Machine learning to predict and prevent rheumatoid arthritis

Principal investigator:

Professor Elina Hyppönen

University of South Australia, Unit of Clinical & Health Sciences

South Australian Health & Medical Research Institute (SAHMRI) Level 8

GPO Box 2471, Adelaide SA 5001

t: +61 8 8302 2518 | **e:** elina.hypponen@unisa.edu.au

Research team:

Professor Catherine Hill

Rheumatology Unit, The Queen Elizabeth Hospital, Adelaide, South Australia, Australia.

Discipline of Medicine, University of Adelaide, Adelaide, South Australia, Australia.

Dr Kitty Pham, Dr Anwar Mulugeta Gebremichael, Dr Iqbal Madakkatel, and Dr Amanda Lumsden

Australian Centre for Precision Health, Clinical & Health Sciences, University of South Australia, Adelaide, Australia.

Dr Kitty Pham was funded by Arthritis Australia to conduct this work.

Consumer advisors:

Ms Sarah Workman, Ms Ruth Lee

Acknowledgement:

This research was conducted with support from Arthritis Australia and funded by ARA Research Fund.

Lay Summary:

Rheumatoid arthritis (RA) is a debilitating disorder, where the body's own defences (the immune system) attack the tissues lining the joints. It is not currently possible to prevent RA as we don't know what causes it, however, we do know that both our genes and some lifestyle factors can increase or lower the risk. To address these gaps in knowledge, we conducted a study at the University of South Australia which used artificial intelligence to discover new risk factors for RA and predict whether these factors increase or lower the risk of disease.

We looked at information from nearly 450,000 individuals who did not have RA or any related diseases at the start of the study. We had access to detailed information on their health, genes, lifestyles, diets, environments, blood markers, medical history, physical measurements and sex-specific factors. The information, comprising around 3,000 characteristics, was fed into a machine learning model (a form of artificial intelligence) that determined which characteristics were most important for predicting who will develop RA within the next 5-10 years.

The model suggested 200 characteristics that were potentially important for predicting RA; 113 of which were confirmed as potential risk factors of RA in further analyses. We learnt that characteristics related to low socioeconomic status, indicators of poor general health, physical inactivity, and history of other joint disorders or lung disease may increase the risk of RA. We also discovered that blood markers involved in inflammation and altered function of the liver and kidneys are important in predicting RA, which presents the opportunity to develop and refine blood tests that check for risk of RA. Our study highlights the need for earlier diagnosis of RA, as our findings showed that individuals experienced symptoms of RA several years before their formal diagnosis.

Next, we looked closely at the genetic risk factors of RA. We identified 24 risk factors that could be modified by lifestyle changes and investigated whether action on these risk factors can reduce risk for those that are genetically predisposed to RA. We found that rheumatoid factor may be a stronger predictor of RA in those who are at high genetic risk of RA, however the effect of other biomarkers and lifestyle factors did not differ notably by genetic susceptibility. This suggests that the key modifiable risk factors of RA apply regardless of genetic susceptibility.

This is the first and largest study of its kind, exploiting large-scale data and newly developed machine learning methods. The findings have added to our understanding of what increases and decreases risk of RA, and will support efforts to improve screening and prevention strategies for RA. Our study was conducted in a UK population, which shares many similarities with the Australian population and will help to guide current guidelines and future research projects in Australia.

Background and Aims of the Project:

Rheumatoid arthritis (RA) is an inflammatory autoimmune disease where the body's immune system attacks the tissues lining of the joints, causing pain, inflammation, and eventually joint damage [1]. The cause of RA is still unknown and careful characterization of risk factors will help with risk stratification and disease prevention.

The aims of this study were to:

1) Identify novel factors associating with the risk of RA, and investigate whether they may increase susceptibility to, or protect from, RA.
2) Explore whether genetic susceptibility to RA modifies these associations to inform on whether approaches may help to mitigate related risks in the presence of genetic susceptibility.

Methods:

This study consisted of secondary data analyses using information from a large UK based cohort study (UK Biobank), implementing approaches from machine learning, together with observational and genetic epidemiological analyses. All associations were investigated among individuals who did not have RA at the start of the study and who were followed up for newly diagnosed RA through record linkage with hospital and primary care data (i.e. "incident cases").

In the first study, we conducted a hypothesis-free data-driven machine learning analyses using information on 445,515 UK Biobank participants [2], including 4,510 newly diagnosed RA cases (8.4 median years follow-up, IQR 5.3 to 10.9). Our initial machine learning model [3] included 2,898 possible predictors of RA (all measured at baseline), of which the model identified 200 as potentially important contributors of risk ($p<0.01$). The identified factors were taken forward to conventional logistic regression analyses to examine the direction of the association and to examine whether their association with RA was independent from social and demographic characteristics (confounders) and known RA risk factors. In these analyses we adjusted for age, sex, assessment centre, ethnicity, Townsend deprivation index (reflecting area deprivation), education, smoking status, body mass index, and physical activity. Analyses involving blood biomarkers were also adjusted for fasting time and aliquot number to account for related influences on the observed concentrations. To reduce the risk of chance findings which may arise from the multiple tests conducted, we used a Bonferroni corrected p-value of $p<2.7\text{x}10^{-4}$ to indicate the significance of association. To describe the associations between blood biomarkers and RA risk we categorised the levels in the quartiles, with Q1 corresponding to the participants in the lowest 20% of the concentrations and Q5 participants in the highest 20%. For exposures that had evidence of sex-interaction indicative of differential associations in men vs. women, we conducted sex-stratified analyses. Finally, to assess the role of reverse causation (i.e. the role of undiagnosed RA explaining the observed association), we also repeated analyses allowing for a 4-year lag time between baseline assessment and diagnosis of RA.

For the second study, we conducted gene-environment interaction analyses to investigate whether the risk factors have a similar role regardless of underlying genetic risk of an individual. We used a polygenic risk score (PRS) [4] to group the participants according to their genetic risk. We first assessed the association between genetic risk and RA in the UK Biobank dataset, and then fitted multiplicative interaction models (using the interaction term exposure x PRS) for 24 potentially actionable risk factors of RA risk. In genetic analyses, we adjusted for age, sex, UK Biobank assessment centre, and genetic principal components and used FDR correction to account for multiple testing.

Project Outcomes:

The findings of our study suggest that a range of lifestyle, psychosocial, and biomarker features are important in predicting risk of RA. From the 200 important features identified by the machine learning pipeline, 113 exposures were associated with RA in the fully adjusted logistic regression models after multiple testing correction.

In addition to the known risk factors of RA, such as older age, female sex, smoking, and physical inactivity, we confirmed that RA risk is associated with indicators of low socioeconomic status and poor health. Interestingly, psychosocial factors representing feelings of anxiety, lethargy, disinterest, anxiety, depression, and neuroticism (highest vs lowest neuroticism score OR 1.44, 95% CI 1.24 to 1.67), and medications or diseases linked to management of RA symptoms were predictive of RA, even if these were measured ~5-10 years before clinical diagnosis. While all these associations persisted in lag time analyses, all were somewhat weakened, suggesting that patients may be experiencing and responding to symptoms of RA many years before they are formally diagnosed. Of disease conditions, history of joint disorder, osteoporosis, hypothyroidism, diverticular disease, and emphysema/chronic bronchitis were associated with more RA, with each condition associating with 48% to 230% higher odds of RA. Markers of inflammation (e.g. C-reactive protein Q5 vs Q1 OR 1.92, 95% CI 1.71 to 2.16), altered liver and kidney function (e.g. cystatin C Q5 vs Q1 OR 1.55, 95% CI 1.38 to 1.74), and a range of blood cell markers (e.g. WBC Q5 vs Q1 OR 1.32, 95% CI 1.19 to 1.46) were also found to associate with the risk of RA, suggesting that blood biomarkers may reflect RA risk several years before the disease diagnoses.

In gene-environment interaction analyses, we show that people at the highest 5% of genetic risk of RA had 3.4 times higher odds of RA and people at top 20% 2.2 times higher odds of RA compared to individuals in to bottom 50% of genetic risk. This shows that it is possible to identify notable risks of RA by screening for common genetic variants. Multiplicative gene-environment interaction analyses found variations in the observed odds of RA by rheumatoid factor ($p_{interaction}$=5.0x10^{-8}), suggesting that rheumatoid factor is a stronger predictive factor for RA in those with high genetic susceptibility. We also investigated additive interactions, with these analyses showing that combining the risk burden of certain risk factors with genetic susceptibility, can notably elevate the risk of disease for an individual.

Impact and Translation:

This is the first and largest study of its kind to explore the risks of RA, using a detailed assessment of individual exposures collected several years prior to RA diagnosis. Indeed, our analyses suggest that people may experience RA related changes in their health and wellbeing many years before the disease is diagnosed. We confirmed the differences in RA risk by various lifestyle and sociodemographic factors, and also, that these associations appear to be relevant regardless of the genetic RA risk of the individual. Combining information on rheumatoid factor with markers of genetic risk enhances the ability to predict those individuals who are likely to receive a RA diagnosis. Also, some other blood biomarkers were associated with RA up to 10 years before the diagnoses, suggesting an opportunity for improved screening strategies and earlier disease detection.

Consumer Involvement:

We recruited two patient research partners as members of our Project Steering Committee, via promotion in the Hospital Research Foundation Group e-newsletter, the Arthritis Australia Advisory Panel and the Arthritis Australia Champions network. The two patient research partners both had lived experience of rheumatoid arthritis and prior experience on research and/or consumer panels. The patient research partners were involved in Committee meetings over Zoom and assisted in setting project priorities, refining analysis models, reviewing study findings, optimising presentation of results, and drafting the lay summary.

Research Outputs:

This project has been presented at the Australian Rheumatology Association 2025 Annual Scientific Meeting, Adelaide, SA, Australia. In addition, we have prepared two scientific publications.

- Pham, K., Madakkatel, I., Mulugeta, A., Lumsden, A., Hill, C., & Hyppönen, E. (2025, May 3-6). Machine learning to predict and prevent rheumatoid arthritis [Poster Presentation]. Australian Rheumatology Association 2025 Annual Scientific Meeting, Adelaide, SA, Australia.
- Pham, K., Madakkatel, I., Mulugeta, A., Lumsden, A., Hill, C., & Hyppönen, E. (2025). Machine learning to discover novel predictors of rheumatoid arthritis. (*Prefinal).*
- Pham, K., Madakkatel, I., Mulugeta, A., Lumsden, A., Hill, C., & Hyppönen, E. (2025). Influence of genetic risk on modifiable risk factors of rheumatoid arthritis. (*In preparation).*

Use of Funding:

The funding was used as described in the original project grant application. It covered the costs of employing a research associate (Dr Kitty Pham) at 0.8 FTE over 5 months (salary and on-costs), and the costs of consumer involvement (honorarium paid per meeting).

References:

1. Malm, K., Bergman, S., Andersson, M. L., Bremander, A., & Larsson, I. (2017). Quality of life in patients with established rheumatoid arthritis: A phenomenographic study. SAGE open medicine, 5, 2050312117713647. https://doi.org/10.1177/2050312117713647

2. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine, 12(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

3. Madakkatel, I., & Hyppönen, E. (2024). LLpowershap: logistic loss-based automated Shapley values feature selection method. BMC medical research methodology, 24(1), 247. https://doi.org/10.1186/s12874-024-02370-8

4. Thompson, D. J., Wells, D., Selzam, S., Peneva, I., Moore, R., Sharp, K., Tarran, W. A., Beard, E. J., Riveros-Mckay, F., Giner-Delgado, C., Palmer, D., Seth, P., Harrison, J., Futema, M., Genomics England Research Consortium, McVean, G., Plagnol, V., Donnelly, P., & Weale, M. E. (2024). A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. PloS one, 19(9), e0307270. https://doi.org/10.1371/journal.pone.0307270